PESQUIMAT , Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. XV N°2, pp. 15-28, Lima - Perú, Marzo 2013

IDENTIFICACIÓN DE OBSERVACIONES INFLUYENTES EN LA DISCRIMINACIÓN DE MUESTRAS DEL GÉNERO MINTHOSTACHYS DE CAJATAMBO Y UNCHOS

Doris Gómez[1], *Olga Solano*[1], *Víctor Osorio*[1], *Liliana Huamán*[1], *Gregoria Ramón* [1], *Jorge Condado*[1], *María Gallardo* [1], *Yakov Quinteros* [2]

Resumen: En el presente estudio se exploraron los métodos para identificar observaciones influyentes en el contexto del análisis discriminante, para clasificar 160 muestras del género *Minthostachys*, 100 recolectadas en la provincia alto andina de Cajatambo, departamento de Lima, y 60 muestras en Unchos, departamento de Ancash, Perú. Los datos utilizados en este estudio provienen de estudios florísticos realizado en los año 2005 y 2006 respectivamente. Las variables morfológicas estudiadas en la rama principal de las *Minthostachys* fueron: longitud del peciolo, longitud y ancho de la hoja. Los estudios taxonómicos y sistemáticos de las muestras se realizaron utilizando el sistema de clasificación de Cronquist, que clasificó para Cajatambo, 51 plantas de *Minthostachys* con abundante pubescencia y 49 con escasa pubescencia; mientras que para Unchos clasificó 40 con abundante pubescencia y 20 con escasa pubescencia. Para las 160 plantas de *Minthostachys*, eliminando una a la vez, se calculó el valor de la Distancia de Mahalanobis, la probabilidad de mala clasificación y las puntuaciones de la función discriminante Fisher (Campbell, 1978; Fung, 1992, 1995). El análisis discriminó correctamente 151 plantas de Minthostachys es decir, el 94,4 % de un total de 160, un valor lo suficientemente grande que mostró la eficacia del análisis discriminante. De las comparaciones de los valores de la Distancia de Mahalanobis, las probabilidades de mala clasificación, las puntuaciones de la función discriminante de Fisher, con y sin la observación evaluada, los mayores cambios en los valores de dichas medidas ocurrieron con las observaciones 110, 112 y 114 , así que hay evidencia de que estas observaciones son influyentes.

Palabras clave: Medidas de influencia; análisis discriminante lineal; género *Minthostachys*.

Abstract: This paper explores the possibility of identifying influential observations in discriminant analysis framework, 160 botanical specimens of the genus *Minthostachys*, pubescent and pubescent not collected in the province of Cajatambo department of Lima and Unchos departament of Ancash. The evaluation of morphological variables in the main branch of each *Minthostachys* being studied were: length of petiole, leaf length and width of the blade. Taxonomic and systematic studies of the samples were performed at the Laboratory of Ethnobotany and Economic Botany of the Natural History Museum and the determination of the species are held in the herbarium of the San Marcos University, using the Cronquist classification system, for Cajatambo which marked 51 plants such as non-pubescent and pubescent 49; while for Unchos ranked 40 with abundant pubescence and 20 escaca pubescence. For the 160 *Minthostachys* plants, eliminating one at a time, we calculated the value of the Mahalanobis distance, the probability of misclassification and the scores of the Fisher discriminant function (Campbell, 1978; Fung, 1992, 1995). The analysis correctly discriminated 151 plants *Minthostachys* ie, 94.4 % of a total of 160, a value large enough to show the effectiveness of discriminant analysis. For the full sample and removing each time one of the samples or observations, we calculated the value of

[1]Facultad de Ciencias Matemáticas, UNMSM
[2]Museo de Historia Natural, UNMSM, e-mail:Yakov281 @hotmail.com

the Mahalanobis Distance, the probability of misclassification, the weightings and scores of discriminant function of Fisher. Comparison of the values of the estimates, with and without the observation under evaluation, it was concluded that observations 110, 112 and 114 were identified as influential.

Key words: Influential observation; linear discriminant analysis; Gender *Minthostachys*.

1. Introducción

Desde 1985, un grupo de investigadores del Departamento de Etnobotánica y Botánica Económica del Museo de Historia Natural de la Universidad Nacional Mayor de San Marcos (UNMSM), tiene interés en los estudios de las poblaciones de plantas medicinales andinas, en particular del género *Minthostachys*, considerada como una de las plantas medicinales más importantes de los Andes del Perú. Es una planta perenne. Cuando joven es herbácea y en la fase adulta es arbusta y puede alcanzar de 1 a 1,5 metros de altura. Sus hojas son verdes, pecioladas, lanceoladas-elípticas y aromáticas. Geográficamente se distribuye a lo largo de los Andes, desde Venezuela, Colombia, hasta la Argentina, creciendo entre los 500 y 4000 metros sobre el nivel del mar. Por lo general crecen en los bordes de los campos o los humedales y desde tiempos inmemoriales es utilizada por los habitantes de los Andes del Perú con fines medicinales, alimenticios y, en los últimos años el aceite extraído de la planta se ha comercializado, por ejemplo, como un repelente de insectos.

Este conjunto de propiedades de la planta se manifiesta como un recurso valioso que podría ser mejor explotado de manera sostenible y podría contribuir a mejorar la salud de los habitantes de los Andes del Perú. En este contexto, fue relevante investigar su potencial, especialmente en Cajatambo, teniendo en cuenta que hasta el año 2004 no se tenía registros de *Minthostachys* de la Provincia de Cajatambo. La provincia de Cajatambo que está ubicada en los Andes occidentales del departamento de Lima, a una altitud de 3376 metros sobre el nivel del mar, con una población de unos 9.618 habitantes, de los cuales 56 % es indígena (INEI,2005).

En 2005 y 2006, un equipo de investigadores del Laboratorio de Etnobotánica del Museo de Historia Natural de la UNMSM hizo un inventario florístico en Cajatambo y Unchos; en las determinaciones taxonómicas, la mayoría de *Minthostachys* fueron identificadas como de la especie *tomentosa*.

El análisis estadístico de las variables morfológicas: longitud del peciolo, ancho del peciolo, el número de venas del cáliz, la longitud de la corola, ancho de la corola, mediante el análisis de componentes principales, mostró dos posibles tipos de *Minthostachys* (Gómez et al, 2008), *Minthostachys* con abundante pubescencia y *Minthostachys* de escasa pubescencia.

Después de muchos años de confusión taxonómica y la indeterminación de sus especies, Schmidt (2008), presentó un resumen general del estado de conocimiento sobre *Minthostachys*, con énfasis en la etnobotánica y el contenido farmacológico del aceite.

En el contexto descrito, el objetivo de este estudio es identificar las observaciones influyentes usando las medidas desarrolladas por Campbell (1978), Fung(1992, 1995) sobre los datos de muestras del género *Minthostachys tomentosa*, colectadas en la provincia de Cajatambo, departamento de Lima y Unchos, departamento de Ancash.

Se trata de una aplicación del método estadístico multivariante conocido en la literatura como el análisis discriminante o discriminación y clasificación que con frecuencia se utiliza para simplificar el tamaño del problema estadístico (Anderson, 1984; Manly, 2005), donde los resultados, pueden verse afectados por la presencia de algunas observaciones que se comportan de manera diferente a la mayoría de los datos, o por la presencia de observaciones discordantes, valores atípicos o influyentes (Beckman y Cook, 1983).

Se han desarrollado muchos estudios con métodos estadísticos o medidas para detectar datos influyentes(Muñoz et al., 2001). Al omitirse en el análisis una observación influyente, éste da lugar a cambios en la estimación de algunos o todos los parámetros que intervienen en el estudio. Se puede considerar como un caso especial de las observaciones discordantes. Una observación es discordante, cuando a juicio del investigador, se encuentra lejos de otras observaciones que forman parte del conjunto de datos analizados(Beckman y Cook, 1983).

Es importante mencionar que habrán observaciones discordantes que no son influyentes, donde las

estimaciones de los parámetros permanecen esencialmente sin cambios cuando en el análisis se omiten dichas observaciones(Beckman y Cook, 1983).
Análisis de influencia ha sido ampliamente estudiado y existen muchas publicaciones en el análisis de regresión (Belsley et al., 1980) y en el contexto del análisis discriminante, el tema fue abordado inicialmente por Campbell (1978), quien propuso las medidas de influencia basadas en la función de influencia dada por Hampel (1974).
Años más tarde, Fung (1992, 1995) en base a la relación que existe entre los coeficientes de la función lineal discriminante de Fisher y los coeficientes del modelo de regresión lineal múltiple, propuso algunas medidas para el análisis discriminante siguiendo la metodología utilizada en el análisis de regresión.
A continuación se presenta la teoría más importante para identificar observaciones influyentes en el contexto del análisis discriminante lineal.

2. Metodología

En el análisis discriminante el objetivo principal es clasificar un individuo $x = (x_1, \cdots, x_p)$ con p medidas, en uno de los k grupos o poblaciones pre determinadas.

2.1 Análisis discriminante lineal en dos grupos

Sean G_1 y G_2 las dos poblaciones o clases de objetos y $X^{(k)} = \left(X_1^{(k)}, \cdots, X_p^{(k)}\right)'$, $k = 1, 2$, un vector aleatorio con valores en $\mathbb{R}^p$, contiene las mediciones de los individuos de cada una de las poblaciones con parámetros, vector de medias y matriz de covarianzas, $\mu^{(k)} = \left(\mu_1^{(k)}, \cdots, \mu_p^{(k)}\right)'$ y Σ_k , y si los valores observados del vector aleatorio $X^{(k)}$, difieren de un grupo a otro a través de sus medidas se construye una regla para clasificar un nuevo individuo, $x = (x_1, \cdots, x_p)'$ de $\mathbb{R}^p$, en una de las dos poblaciones grupos G_1, G_2.
Dadas las consideraciones anteriores, se toman muestras aleatorias de cada una de las poblaciones,para estimar los parámetros de interés, donde $\overline{x}^{(1)}$, $\overline{x}^{(2)}$ y S son las estimaciones de los vectores de medias y de la matriz de covarianzas común $\Sigma = \Sigma_k$, respectivamente.
Fisher (1936), propuso encontrar una combinación lineal del vector x, $Y = \widehat{\alpha}'x$, en cada población, de manera que sea máxima la razón de la diferencia de medias al cuadrado de las combinaciones lineales, respecto a su varianza, es decir, que sea máximo

$$\lambda = \frac{\left(\widehat{\alpha}'\overline{x}^{(1)} - \widehat{\alpha}'\overline{x}^{(2)}\right)^2}{\widehat{\alpha}' S \widehat{\alpha}}$$

Se demuestra que el vector $\widehat{\alpha}$ es proporcional a $S^{-1}\left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)$, y la combinación lineal,

$$Y = \left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)' S^{-1} x \tag{1}$$

se conoce como la función lineal discriminante de Fisher.
Haciendo,

$$\widehat{\alpha} = S^{-1}\left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)$$

se define la siguiente regla de clasificación:

Asignar x al grupo G_1 si

$$\widehat{\alpha}' x - \tfrac{1}{2}\left(\overline{x}^{(1)} + \overline{x}^{(2)}\right) \geq 0\,, \tag{2}$$

caso contrario asignar al grupo G_2

Algunos aspectos importantes relacionados con el tema de discriminación en dos grupos son:

a) La Distancia de Mahalanobis poblacional, $\Delta^2 = \left(\mu^{(1)} - \mu^{(2)}\right)' \Sigma^{-1} \left(\mu^{(1)} - \mu^{(2)}\right)$, estimada por la expresión

$$\widehat{\Delta}^2 = D^2 = \left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)' S^{-1} \left(\overline{x}^{(1)} - \overline{x}^{(2)}\right) \tag{3}$$

b) La probabilidad de clasificar equivocadamente (probabilidad de mala clasificación) una observación del grupo G_j en el grupo G_i, según una regla de clasificación, digamos la regla R, dada por

$$P(i/j; R) = \phi\left(-\tfrac{1}{2}D^2\right) \tag{4}$$

donde ϕ es la función de distribución normal acumulada en el punto $\left(-\frac{1}{2}D^2\right)$, $i, j = 1, 2 \;\; i \neq j$.

c) La función lineal discriminante dada por Fisher

$$Y = \left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)' S^{-1} x \tag{5}$$

d) Los escores o puntuaciones de la función lineal discriminante de Fisher dados por

$$\widehat{\alpha}' x - \tfrac{1}{2}\widehat{\alpha}' \left(\overline{x}^{(1)} + \overline{x}^{(2)}\right) \tag{6}$$

Uno de los problemas relacionados con el análisis discriminante, es la presencia de observaciones influyentes que afectan los valores de las medidas: la Distancia de Mahalanobis, la probabilidad de mala clasificación, los escores o puntuaciones de la función discriminante y la función lineal discriminante. Frente a este problema, para detectar las observaciones influyentes, se ha desarrollado el análisis de influencia. La idea básica detrás del análisis de influencia es comparar las estimaciones de las medidas: la Distancia de Mahalanobis, la probabilidad de mala clasificación, la función lineal discriminante de Fisher y las puntuaciones de la función lineal discriminante, con y sin la observación considerada influyente.
En diversos estudios sobre el tema, el tipo de perturbación mas usada para evaluar la influencia de una observación, es la omisión de la observación supuestamente influyente, en la estimación de las medidas involucradas en el análisis discriminante(Muñoz et al, 2001). Por eso es importante evaluar el efecto de la i - ésima observación multivariante, $x_i = (x_{i1}, \cdots, x_{ip})'$, en cada una de las estadísticas involucradas en el análisis discriminante.

2.2. Medida de influencia para la Distancia de Mahalanobis

Para evaluar la posible influencia de la observación multivariante, x en la Distancia de Mahalanobis de la muestra $D^2 = \left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)' S^{-1} \left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)$, Fung(1992) propuso la siguiente función de influencia

$$\widehat{I}M\left(x; \widehat{\Delta}^2\right) = w_1 \left[\widehat{\Psi} - w_1{}^{-1}\right]^2 \tag{7}$$

donde: $\widehat{\Psi} = \widehat{\alpha}'\left(x - \overline{x}^{(k)}\right)$, y w_k es el peso de cada grupo en la formación de la matriz de covarianzas, $w_1 = \dfrac{(n_1 - 1)}{n_1 + n_2 - 2}$ y $w_2 = 1 - w_1$.

Esta medida depende básicamente de la estadística $\widehat{\Psi}$, que compara cada observación con el vector de mediciones del grupo al que pertenece, ponderado por el coeficiente de la función discriminante de Fisher.

2.3. Medida de influencia para la probabilidad de mala clasificación

La probabilidad de mala clasificación cuantifica la probabilidad de ubicar erroneamente el vector con medidas $x = (x_1, \cdots, x_p)'$, en el grupo G_i cuando en realidad pertenece al grupo G_j. Para una regla de clasificación R, la probabilidad de clasificar equivocadamente una observación fue definida en (4) como $P(i/j; R) = \phi\left(-\frac{1}{2}D^2\right)$.

Para evaluar la posible influencia de la i-ésima observación multivariante sobre la probabilidad de mala clasificación, Hampel (1974) propuso la siguiente función de influencia

$$\widehat{I}(x, MP) = (n_1 - 1)\left[\phi\left(-\tfrac{1}{2}D\right) - \phi\left(-\tfrac{1}{2}D_{(i)}\right)\right]^2 \tag{8}$$

donde:

D: es la raíz cuadrada de la Distancia de Mahalanobis con todos los datos o la muestra total, y

$D_{(i)}$: es la raíz cuadrada de la Distancia de Mahalanobis omitiendo la i-ésima observación.

Si las estimaciones de los vectores de medias, omitiendo la i-ésima observación del grupo k son, $\overline{x}^k_{(i)}$, la función discriminante lineal es $Y = \left(\overline{x}^{(1)}_{(i)} - \overline{x}^{(2)}\right) S^{-1}x$.

$$Y = \widehat{\alpha}'_{(i)}x \tag{9}$$

donde $\widehat{\alpha}_{(i)} = \left(\overline{x}^{(1)}_{(i)} - \overline{x}^{(2)}\right) S^{-1}$ son los coeficientes de las función lineal discriminante cuando se ha omitido la i-ésima observación del grupo 1. En este caso, la regla de clasificación, omitiendo la i-ésima observación del grupo 1 es la siguiente

---

Clasificar x al grupo G_1 cuando

$$\widehat{\alpha}'_{(i)}\left[x - \tfrac{1}{2}\widehat{\alpha}'\left(\overline{x}^{(1)}_{(i)} + \overline{x}^{(2)}\right)\right] > 0\,, \tag{10}$$

caso contrario asignar al grupo G_2 .

Fung (1992). propuso la siguiente medida de influencia para evaluar el efecto de la i-ésima observación sobre la probabilidad de mala clasificación

$$DMP_i = \left[\tfrac{1}{2}\left(P^{(1)}_{(i)} + P^{(2)}_{(i)}\right)\right] - \left[\phi\left(-\tfrac{1}{2}D\right)\right] \tag{11}$$

donde:

$$P^{(1)}_{(i)} = \phi\left[\frac{-\widehat{\alpha}'_{(i)}\left(\overline{x}^{(1)} - \overline{x}^{(2)}\right) - \widehat{\alpha}'_{(i)}\left(\overline{x}^{(1)} - \overline{x}^{(1)}_{(i)}\right)}{2G}\right]$$

$G^2 = \widehat{\alpha}'_{(i)}\, S\, \widehat{\alpha}_{(i)}$.

2.4. Medida de influencia para la probabilidad de mala clasificación con la aproximación de Taylor

Considerando la aproximación de segundo orden del polinomio de Taylor, alrededor de $-\frac{1}{2}D$, la propuesta de Fung (1992) es una medida alternativa a la ecuación (11). Así se tiene la medida DMP para la i-ésima observación

$$DMP_i \cong \frac{\phi\left(-\frac{1}{2}\cdot D\right)}{4\cdot D(n_1-1)^2}\left[\left(1 - w_k\cdot\widehat{\Psi}\right)^2\left(d_i^2 - \frac{\widehat{\Psi}_i^2}{D^2}\right)\tfrac{1}{4}\widehat{\Psi}_i^2\right] \tag{12}$$

donde

$$\begin{aligned} d_i^2 &= \left(x_i^{(k)} - \overline{x}^{(k)}\right)' S \left(x_i^{(k)} - \overline{x}^{(k)}\right) \\ \widehat{\Psi}_i &= \left(x_i^{(k)} - \overline{x}^{(k)}\right)^T S \left(x_i^{(k)} - \overline{x}^{(k)}\right) \end{aligned}$$

$x_i^{(k)}$ es la i-ésima observación del grupo k, donde $k = 1, 2$.

2.5. Medida de influencia para los escores de la función discriminante

Fung (1995) propuso una medida para evaluar la influencia de una observación sobre los escores o puntuaciones de la función discriminante de Fisher, siguiendo la metodología propuesta por Cook y Weisberg (1982), basado en la cuantificación del efecto de la omisión de una observación sobre el vector de parámetros, teniendo en cuenta la relación de equivalencia entre los coeficientes de la función discriminante de Fisher y los coeficientes del modelo de regresión lineal múltiple de Anderson (1984), donde

$\widehat{\alpha}'x - \frac{1}{2}\widehat{\alpha}'\left(\overline{x}^{(1)} + \overline{x}^{(2)}\right)$, son las puntuaciones de la función lineal discriminante, que también pueden ser denotados como $\beta'x$, donde

$$\beta' = \left[-\frac{1}{2}\widehat{\alpha}'\left(\overline{x}^{(1)} + \overline{x}^{(2)}\right), \widehat{\alpha}'\right]$$

$$x' = [1, x']$$

$\beta_{(i)}$ es el vector β, sin la i-ésima observación del grupo 1.

El efecto de la i-ésima observación se evalúa a través de la diferencia de los escores de la función discriminante, con y sin esta observación, o sea, mediante la diferencia

$$\beta'x - \beta'_{(i)}x.$$

Fung (1995) propuso la siguiente medida:

$$E2 = t \cdot \widehat{\beta}_1^2 + (1-t)\,\widehat{\beta}_2^2 + V \tag{13}$$

donde: $t = \frac{n_1}{n}$,

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\left(\widehat{\alpha} - \widehat{\alpha}_{(i)}\right)^T \left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)}{2} - \frac{\widehat{\alpha}_{(i)}\left(\overline{x}^{(1)} - \overline{x}^{(1)}_{(i)}\right)}{2} \\
\widehat{\beta}_2 &= \frac{-\left(\widehat{\alpha} - \widehat{\alpha}_{(i)}\right)^T \left(\overline{x}^{(1)} - \overline{x}^{(2)}\right)}{2} - \frac{\widehat{\alpha}_{(i)}\left(\overline{x}^{(1)} - \overline{x}^{(1)}_{(i)}\right)}{2} \\
V &= \left(\widehat{\alpha} - \widehat{\alpha}_{(i)}\right)^T S \left(\widehat{\alpha} - \widehat{\alpha}_{(i)}\right)
\end{aligned}$$

3. Resultados y discusión

Para el presente trabajo se usaron los datos de 160 muestras de *Minthostachys* tomentosa, recolectadas en la Provincia de Cajatambo, Departamento de Lima en el año 2005 y en Unchos, Departamento de Ancash en el año 2006 (Figura 1)

Figura 1: Hojas de *Minthostachys*

Los muestreos se realizaron entre los meses de enero y junio de los años 2005 y 2006, a una altitud de 2.800 a 3.600m en las comunidades indígenas de la provincia de Cajatambo y Unchos. Los datos fueron recolectados durante la temporada de lluvias de enero a marzo y durante los meses secos, de abril a junio. Las muestras fueron inventariadas en el Laboratorio de Etnobotánica y Botánica Económica del Museo de Historia Natural de la UNMSM, según el catálogo de angiospermas y gimnospermas del Perú (Brako; Zarucchi, 1993), que muestra la siguiente distribución del género Minthostachys en el Perú. (Tabla 1) y en cuyas instalaciones se hicieron los estudios de taxonomía y sistemática de las muestras por el sistema de clasificación de Cronquist.

Tabla 1. Distribución de las especies de *Minthostachy* en el Perú (Brako; Zarucchi, 1993)

Especies	**Altitud** (m.s.n.m)	**Ubicación Geográfica**
Minthostachys glabrescens (Bentham)	2500 - 4000	Apurimac, Cajamarca, Cuzco, Junín.
Minthostachys mollis(*Grisebach*)	500 - 3500	Amazonas, Arequipa, Cajamarca, Cuzco, Huánuco, Junín, Lima, La Libertad, Piura.
Minthostachys setosa (Briquet) Epling	1000 - 1500	Puno
Minthostachys tomentosa (Bentham)	2000 - 3500	Amazonas, Cajamarca, Cuzco, Huánuco, Junín, Lima, La Libertad, Ancash.
Minthostachys andina (Britton) Epling	2000 - 2500	Cuzco
Minthostachys mandoniana (Briquet) Epling	1000 - 1500	Ayacucho
Minthostachys salicifolia Epling	2500 - 3000	Ayacucho

Se recolectaron 100 muestras en Cajatambo y 60 en Unchos.

Para realizar el análisis discriminante se consideró las siguientes variables: $X_1 = Longitud\ del\ peciolo$ (cm); $X_2 = Longitud\ de\ la\ hoja$ (cm) y $X_3 = Ancho\ de\ la\ hoja$ (cm), conforme se ilustra en la Figura 2.

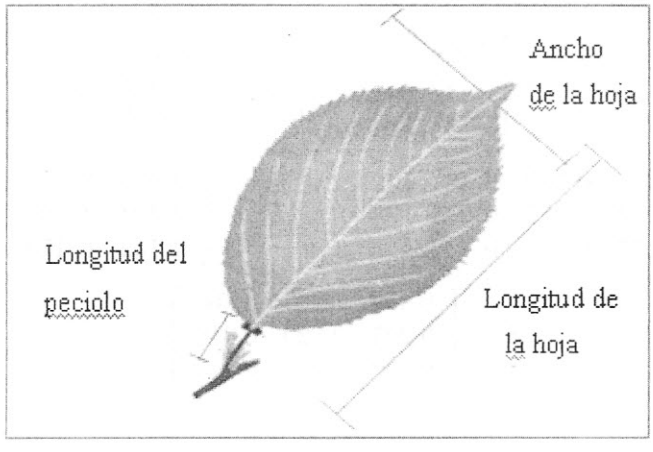

Figura 2: Partes de la hoja de *Minthostachys*

Los datos fueron procesados con el software estadístico SPSS, Statistical Package for the Social Sciences versión 17 y el Matlab, Versión 7.1, adoptándose el nivel de significación 0,05.

La Tabla 2 muestra las estadísticas descriptivas y el análisis de varianza para cada una de las variables univariantes. Los valores de las estadísticas F y las probabilidades asociadas (p-valor) permiten rechazar la hipótesis de igualdad de medias de cada variable al nivel de significación de 0,05.

Tabla 2. Media aritmética y desviación estándar de las variables y resultado del test de igualdad de medias para cada variable

Variabes (cm)	*Minthostachys*		F(1,98)	p - valor
	Cajatambo	Unchos		
Longitud de peciolo	$0,83 \pm 0,51$	$2,67 \pm 0,79$	320,6	0,000
Longitud de la hoja	$3,45 \pm 0,70$	$4,23 \pm 1,05$	31,7	0,000
Ancho de hoja	$1,94 \pm 0,49$	$2,10 \pm 0,41$	4,0	0,046

La Tabla 3 muestra el valor de Lambda de Wilks, la relación entre la suma de los cuadrados entre grupos y la suma de cuadrados total, la prueba compara los vectores de medias multivariantes o las medias de las funciones discriminantes en ambos grupos. Lambda de Wilks se transforma en una variable que es asintóticamente tiene distribución chi-cuadrado $\left(\chi^2 = \left(n - k - \frac{1}{2}(p - k + 2)\right) \ln(\Lambda)\right)$. Se postuló la hipótesis de que las Minthostachys de Cajatambo y Unchos, proceden de poblaciones con vectores de medias significativamente diferentes, o que las medias de las funciones discriminantes son significativamente diferentes. Observando el valor de la estadística Lambda de Wilks (0,274) o el valor de chi-cuadrado, que se muestra en la Tabla 3, se rechaza la hipótesis de igualdad de los vectores de medias entre *Minthostachys* de Cajatambo y *Minthostachys* de Unchos. Es decir, las diferencias de los vectores de medias son estadísticamente significativas a un nivel de significación de 0,05. O sea, son estadísticamente significativas las diferencias entre los vectores de medias, al nível de significación de 0,05.

Tabla 3. Test, de las funciones discriminantes o de igualdad de vectores de medias multivariantes

Test de la función	Lambda de Wilks	Chi cuadrado	Grados libertad	P - valor
1	0,274	202,367	3	0,000

a continuación se presenta los resultados, los vectores de medias y las matrices de covarianzas siguiendo la notación del analisisís discriminante:

$$\overline{X}^{(1)} = \begin{bmatrix} 0,83 \\ 3,45 \\ 1,94 \end{bmatrix} \qquad \overline{X}^{(2)} = \begin{bmatrix} 2,67 \\ 4,23 \\ 2,10 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 0,275 & 0,212 & 0,177 \\ 0,212 & 0,487 & 0,244 \\ 0,177 & 0,244 & 0,242 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 0,627 & 0,193 & 0,132 \\ 0,193 & 1.103 & 0,170 \\ 0,132 & 0,170 & 0,170 \end{bmatrix}$$

$$S = \begin{bmatrix} 0{,}406 & 0{,}205 & 0{,}160 \\ 0{,}205 & 0{,}720 & 0{,}216 \\ 0{,}160 & 0{,}216 & 0{,}215 \end{bmatrix}$$

donde; $S = \dfrac{99S_1 + 59S_2}{158}$.

El vector de coeficientes de la función lineal discriminante de Fisher,

$$\widehat{\alpha} = S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right) = \begin{bmatrix} -6,14 \\ -0,71 \\ 4,58 \end{bmatrix}$$

donde según la ecuación (1), $Y = -6,14X_1 - 0,71X_2 + 4,58X_3$, es la función discriminante de Fisher.

Según la ecuación (3) tenemos el valor de la Distancia de Mahalanobis igual a 11,14 y la probabilidad de mala classificación según la ecuación (4) alcanzó el valor de 0,048 conforme se muestra en la Tabla 4.

Tabla 4. Valores de las estadísticas con todas las observaciones

Vector de los coeficientes	Distancia de Mahalanobis	Prob. de mala clasificación	% de obs. clasificadas equivocadamente	Observaciones clasificadas equivocadamente
$\begin{bmatrix} -6,14 \\ -0,71 \\ 4,58 \end{bmatrix}$	11.14	0.048	5,6 %	56 64 65 82 109 110 112 113 114

Cada observación o cada una de las 160 muestras de *Minthostachys* fueron evaluadas en la ecuación (6) dando origen a los escores o puntuaciones discriminantes. La muestra de *Minthostachys* de Cajatambo (1-Grupo 1), con los códigos 56, 64, 65 y 82 fueron clasificadas como de Unchos y la muestra de *Minthostachys* de Unchos (2-Grupo 2), con códigos 109, 110, 112, 113 y 114 fueron mal clasificadas como de Cajatambo, lo que representó el 5,6 %(9/160) de observaciones mal clasificadas. La Figura 3 muestra los escores de la función linear discriminante de Fisher para las 160 observaciones.

Tabla 5. Estadísticas de la clasificación

N° de la observación	Grupo verdadero	Clasificado al grupo	Escores discriminantes
1	1	1	-1,62
2	1	1	-1,30
3	1	1	-1,28
4	1	1	-1,78
5	1	1	-1,88
56	1	1	0,44
64	1	1	0,60
65	1	1	0,43
82	1	1	0,68
94	1	1	0,10
110	2	2	-2,26
112	2	1	-1,66
113	2	2	-1,15
114	2	2	-1,53
159	2	1	3,22
160	2	1	2,17

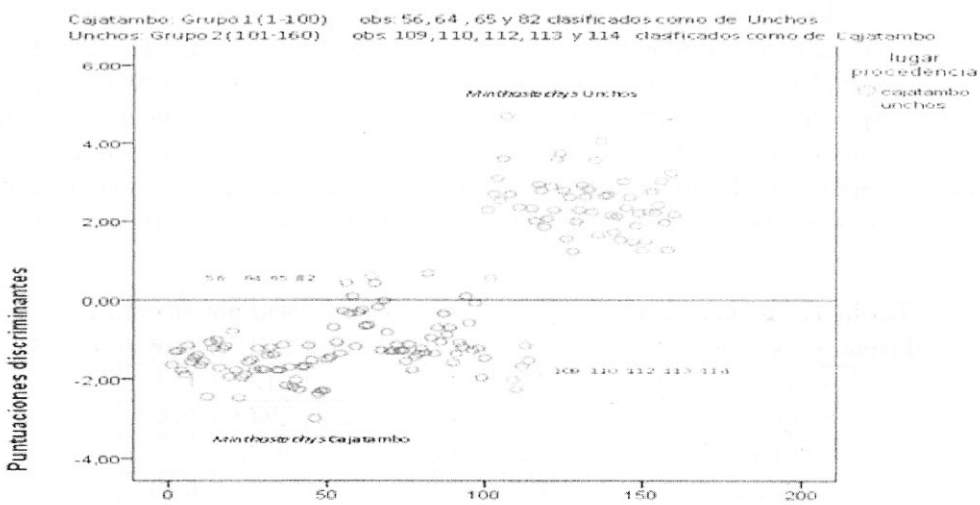

Figura 3: Escores de la función lineal discriminante

En la Tabla 5 se presentan algunos resultados asociados al análisis discriminante con las 160 observaciones de *Minthostachys*. El número asignado a cada elemento de la muestra, el grupo verdadero al cual pertenece la observación de la muestra, el grupo al que fueron asignados los individuos de acuerdo a la ecuación de la clasificación(2) y las puntuaciones o escores discriminantes para cada muestra de *Minthostachys*.
Eliminando cada vez una observación, o sea con 159 observaciones por vez, se encuentra el vector de coeficientes de la función lineal discriminante de Fisher, la Distancia de Mahalanobis y las observaciones mal clasificadas. Este procedimiento se repite, generando para cada una de las repeticiones, los coeficientes de la función discriminante, la Distancia de Mahalanobis, la probabilidad de clasificación errada y las observaciones mal clasificadas. Entre todos los casos, los mayores cambios en las estadísticas relacionadas con el análisis discriminante fueron para las observaciones 107, 109, 110, 112 y 114 como se muestra en la Tabla 5.

Tabla 6. Estimaciones de las medidas relacionadas con el análisis discriminante omitiendo las observaciones: 107 109 110 112 y 114.

Medidas	Omitiendo la observación 107	109	110	112	114
Vector de coeficientes de la función lineal discriminante	$\begin{bmatrix} -6,43 \\ -0,48 \\ 4,68 \end{bmatrix}$	$\begin{bmatrix} -7,04 \\ -0,68 \\ 5,17 \end{bmatrix}$	$\begin{bmatrix} -7,09 \\ 0,91 \\ 5,55 \end{bmatrix}$	$\begin{bmatrix} -6,82 \\ -0,78 \\ 4,76 \end{bmatrix}$	$\begin{bmatrix} -1,51 \\ -0,78 \\ 4,84 \end{bmatrix}$
Distancia de Mahalanobis	11,30	12,95	13,17	12,66	12,54
Probalilidad de mala clasificación	5,0 %	3,8 %	3,8 %	5,0 %	5,0 %
Porcentaje de observaciones mal clasificadas	5,0 %	3,8 %	3,8 %	5,0 %	5,0 %
Observaciones mal clasificadas	56, 64 82, 107, 108, 110, 111, 112	56, 64 108, 110, 111, 112	56, 64, 109, 110, 111, 112	56, 64, 65, 82, 109, 111, 113, 114	56, 64, 65, 82, 109, 110 111, 112

La Tabla 7 muestra los valores de las observaciones identificadas como potencialmente influyente según la medida de influencia evaluada. Los valores más altos para la medida de influencia de la ecuación (7) corresponden a las observaciones 109, 110, 112 y 114; para la medida de influencia de la ecuación (8) corresponden a las observaciones 109, 110, 112 y 114; para la media de influencia de la ecuación (11) corresponden a las observaciones 107, 109, 110 y 112; para la medida de influencia de la ecuación (12) corresponden a las observaciones 107, 109, 110 y 112; para la medida de influencia de la ecuación (13) los valores más grande corresponden a las observaciones, 109, 110, 112 y 114.

Tabla 7. Medidas de observaciones identificadas como potencialmente influyentes según las diferentes medidas de influencia (MI) (Datos actuales)

Medida de influencia	107	109	110	112	114
Ecuación (7)	31,4	146,7	162,5	124,8	116,9
Ecuación (8)	0,28	0,66	0,74	0,56	0,52
Ecuación (11)	0,31	0,15	0,17	0,22	0,14
Ecuación (12)	0,11	0,07	0,08	0,10	0,06
Ecuación (13)	0,06	0,64	0,84	0,44	0,36

Las figuras 4, 5, 6, 7 y 8, muestran las respectivas puntuaciones.

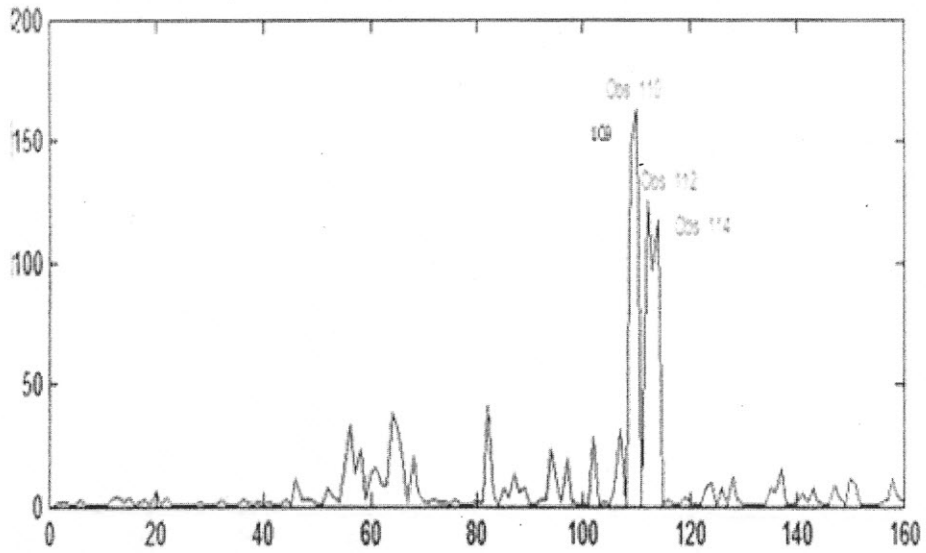

Figura 4: Medida de influencia para la distancia de Mahalanobis

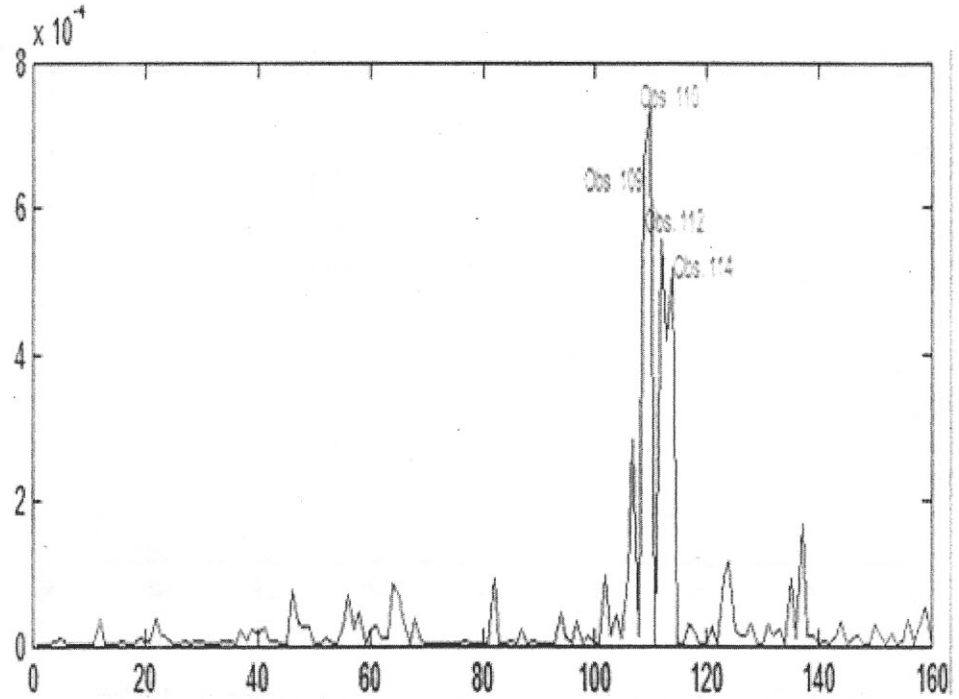

Figura 5: Medida de influencia para probabilidad de mala clasificación

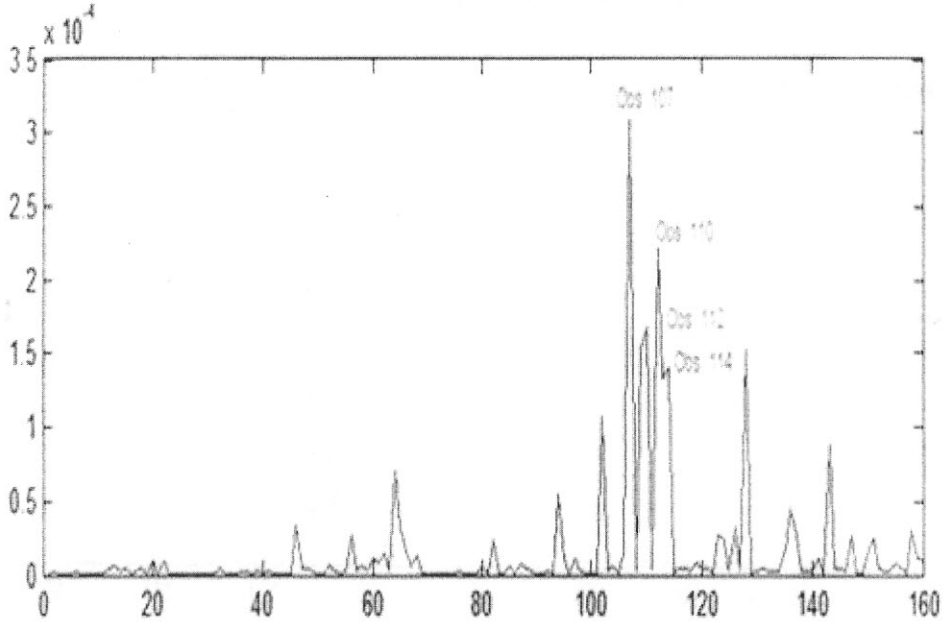

Figura 6: Medida de influencia alternativa para la probabilidad de mala clasificación

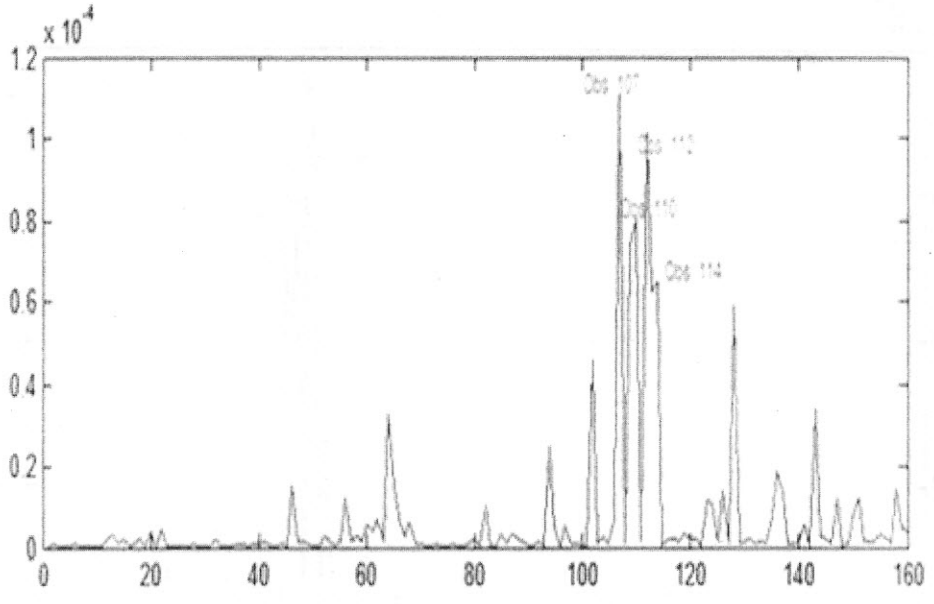

Figura 7: Medida de influencia según la aproximación de Taylor

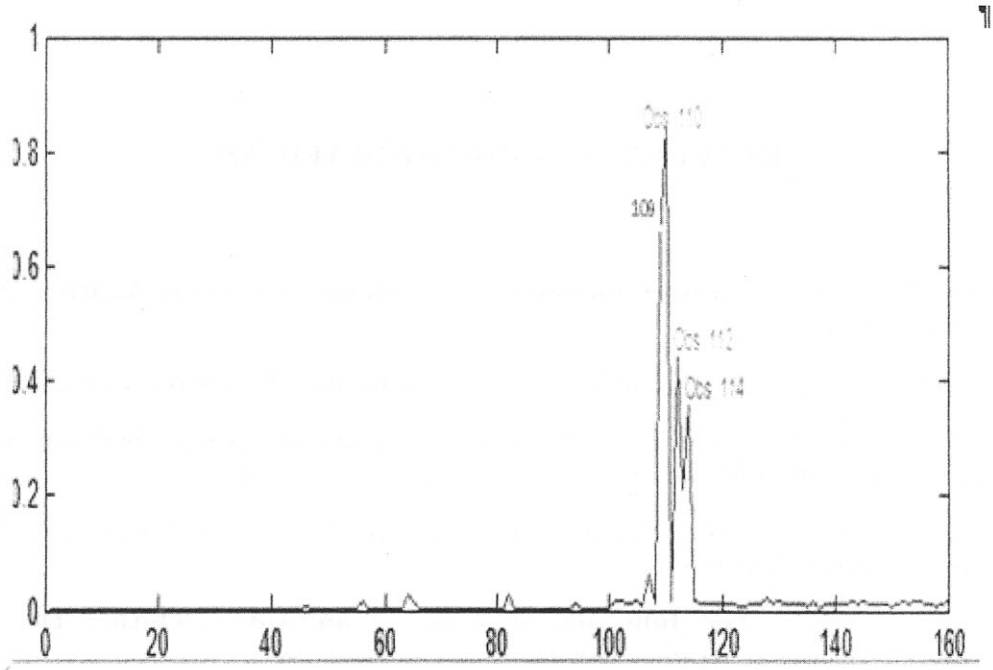

Figura 8: Medida de influencia para los escores de la función lineal discriminante

Los coeficientes de correlación de Pearson entre los valores obtenidos con las diferentes medidas de influencia indican que son estadísticamente significativos al 0,05(p-valor $< 0{,}05$), es decir, todas las medidas, coinciden en identificar a las mismas observaciones como observaciones potencialmente influyentes. Los resultados se muestran en la Tabla 8.

Tabla 8. Medidas de asociación entre las diferentes Medidas de Influencia

Puntuaciones	Coeficientes de Correlación
Puntuaciones de las ecuaciones (7) y (11)	0,48
Puntuaciones de las ecuaciones (7) y (12)	0,76
Puntuaciones de las ecuaciones (7) y (13)	0,68
Puntuaciones de las ecuaciones (11) y (12)	0,84
Puntuaciones de las ecuaciones (11) y (13)	0,46
Puntuaciones de las ecuaciones (12) y (13)	0,69

4. Conclusiones

Al aplicar la metodología del análisis discriminante, las observaciones: 56, 64, 65 y 82, *Minthostachys* de Cajatambo (1-Grupo 1), fueron clasificadas como de Unchos y las observaciones 109, 110, 112, 113 y 114, *Minthostachys* de Unchos (2-Grupo 2), fueron mal clasificadas como de Cajatambo, lo que representó el 5,6 %(9/160) de observaciones mal clasificadas.

Teniendo en cuenta todas las medidas de influencia representadas por las ecuaciones (7), (8), (11), (12) y (13), las observaciones 107, 109, 110 , 112 y 114, fueron identificadas como potencialmente influyentes.

Los mayores cambios en las medidas involucradas en el análisis discriminante, cada vez que se eliminó una observación, ocurrieron cuando se retiraron las observaciones 110, 112 y 114, cuyos valores se presentan en la Tabla 6. Así, puede concluirse que dichas observaciones fueron influyentes.

Los valores de los coeficientes de correlación entre las puntuaciones obtenidas con las diferentes medidas de influencia son significativos al 0,05, (p-valor $< 0{,}05$), Tabla 7, que indica que hay concordancia entre las observaciones identificadas como potencialmente influyentes a través de las diversas medidas de influencia.

REFERENCIAS BIBLIOGRÁFICAS

[1] Anderson, T. W. (1984). **An introduction to Multivariate Statistical Analysis.** 2. ed. New York: Wiley e Sons.

[2] Beckman, R., Cook, R. (1983). **Outlier...s(with discution). Technometrics**, 25(2), 119-149.

[3] Belsley, D., Kuh, E. & Welsch, R. (1980). **Regression diagnostics. Identifying influential data and sources of collinearity**. New York: John Wiley & Sons.

[4] Brako, L.,Zarucchi, J. (1993). **Catálogo de Angiospermas y Gimnospermas del Perú. Missouri Botanical Garden**. USA.

[5] Campbell, N. (1978). **The Influence Function as an Aid in Outlier Detection in Discriminant Analysis. Applied. Statistics**, 27(3), 251-258.

[6] Cook, R.D., Weisberg, S. (1982). **Residual and Influence in Regresión**. New York: Chapman & Hall.

[7] Fisher, R. (1936). **The use of multiple measurements in taxonomic problems. Annals of Eugenics**, 7(2), 179-188.

[8] Fung, W.K. (1992). **Diagnostics in Linear Discriminant analysis. Statistics and Probability Letters**, 13, 279-285.

[9] Fung, W.K. (1995). **Some diagnostic measures in discriminant analysis. J. Am. Stat. Assoc.**, 90, 952-956.

[10] Gómez, D. et al. (2008). **Determinación de patrones de variación morfológica del género Minthostachys en Unchos y Cajatambo mediante métodos estadísticos multivariantes de reducción de datos. Pesquimat. Revista de investigación de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos**, Lima, Perú, XI, (1), 53-66.

[11] Hampell, F., (1974). **Influence curve and its role in robust estimation. J. Am. Stat. Assoc.**, 69, 383-393.

[12] INEI (2005). **Censo de Población y Vivienda**. Instituto Nacional de Estadística, Lima, Perú.

[13] Manly, B. (2005). **Multivariate statistical methods**.3. ed. New York: Chapman & Hall/CRC.

[14] Muñoz, J., Moreno, J., Gómez, T. & Enguix, A. (2001). **El sesgo condicionado en el análisis de influencia: una Revisión**. Facultad de Matemática, Universidad de Sevilla. Questiió, 25, (2), 263-284.

[15] Schmidt, L. (2008). **Ethnobotany, biochemistry and pharmacology of Minthostachys (Lamiaceae)**. J. Ethnopharmacol, 118(3), 343-353.